Eur. Phys. J. B **75**, 357–364 (2010)
DOI: 10.1140/epjb/e2010-00120-8

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Heavy-tailed distribution of cyber-risks

T. Maillart[a] and D. Sornette[b]

Department of Management, Technology and Economics, ETH Zurich, Kreuzplatz 5, 8032 Zurich, Switzerland

Received 7 December 2009 / Received in final form 10 March 2010
Published online 7 April 2010 –

Abstract. With the development of the Internet, new kinds of massive epidemics, distributed attacks, virtual conflicts and criminality have emerged. We present a study of some striking statistical properties of cyber-risks that quantify the distribution and time evolution of information risks on the Internet, to understand their mechanisms, and create opportunities to mitigate, control, predict and insure them at a global scale. First, we report an exceptionnaly stable power-law tail distribution of personal identity losses per event, $\Pr(\text{ID loss} \geq V) \sim 1/V^b$, with $b = 0.7 \pm 0.1$. This result is robust against a surprising strong non-stationary growth of ID losses culminating in July 2006 followed by a more stationary phase. Moreover, this distribution is identical for different types and sizes of targeted organizations. Since $b < 1$, the cumulative number of all losses over all events up to time t increases faster-than-linear with time according to $\simeq \mathbf{t^{1/b}}$, suggesting that privacy, characterized by personal identities, is necessarily becoming more and more insecure. We also show the existence of a size effect, such that the largest possible ID losses per event grow faster-than-linearly as $\sim S^{1.3}$ with the organization size S. The small value $b \simeq 0.7$ of the power law distribution of ID losses is explained by the interplay between Zipf's law and the size effect. We also infer that compromised entities exhibit basically the same probability to incur a small or large loss.

1 Introduction

The Internet has developed into a global system of interconnected computer networks that allows the exchange of data between millions of private and public, academic, business, and government organizations. By making possible new forms of social interactions as well as new ways to probe them, the Internet provides a unique tool for studying the development and the organization of an archetypical complex system.

But, as in all complex biological and social systems known to us, upgrades of capacity, improved networking and additions of functionalities come together with its bundle of parasites, viruses and criminals. We ask what are the laws, in any, codifying these dynamics, and what are the possible roles and consequences of such apparently negative developments?

In biology, there is a growing realization that evolution has been driven and shaped by bacteria and viruses [1]. Similarly, social organizations, which are founded on laws and regulations, and which are anchored on national (as well as sub- and super-national) boundaries, have arguably been shaped in significant part by the need to coordinate and cooperate in the face of disruptions emerging from within and from the outside. In this vein, we ask what may the exploding level of criminality and of unlawful exploitation of the Internet teach us on the organization of other complex systems? Are there robust dynamics or universal laws that can be inferred and tested? What does the fact, that electronic crime has appeared and developed concommittantly with the growth of the Internet, teach us on its organization, its vulnerabilities and its future development?

Given the breadth of these questions, our contribution is to focus on a specific criminality which is now becoming rampant, the theft of personal information (ID thefts). Using the most complete dataset from the Open Security Foundation [2], we are able to identify an explosive growth of ID losses followed by a regime which seems to have matured into a stationary phase. We document a very heavy-tailed power-law distribution (an often reported hallmark of complex systems) of severities of ID theft events, which is robust over all time periods and across different types of social organizations (private and public). By quantifying the scaling of losses as a function of organization sizes, we unearth a significant size effect.

2 Maturation and severity of ID losses: non-stationary and stationary properties

2.1 Contextual data description

From early (gentle) hackers breaking in systems to demonstrate their skills, some turned into seasoned "black hats"

[a] e-mail: thomas.maillart@gmail.com
[b] e-mail: dsornette@ethz.ch

making money as part of an explosively growing business based on ubiquitous Internet insecurity [3,4]. Compared with the attacks that used to disrupt networks on a large scale, most electronic attacks nowadays extract out valuable data while remaining quite furtive [5]. This can be likened to an electronic form of massive parasitism. In terms of monetary value and volume, one of the largest types of data targeted by pirates is personal identity information (ID), such as credit card numbers, social security numbers, banking accounts, and medical files. Since each ID theft or leakage is a "loss of control" of one's individual private data, it can be considered already as a damaging event, forerunning the potential realized financial and/or social losses [6]. Actually, stealing ID's is the goal which is common to a wide spectrum of non-destructive Internet attacks focused on profit, from botnets to highly tailored attacks [7–10]. The (uncontrolled) dissemination of personal information raises the important social issue of people's identity resilience in the information technology era [5,6]. In our quantitative study of cyber-risks, we take a ID theft as a usable elementary unit of cyber-risks, for two main reasons. First, it provides a natural metric of the "permeability" of information systems, guiding towards the identification of the underlying mechanisms. Second, it offers a common basis, or currency, to compare a large variety of heterogeneous events involving many different types of organizations.

ID loss event data have been thoroughly collected by several independant organizations. We use the most complete dataset from the Open Security Foundation [2], that contains 956 documented events reported mainly in the USA between year 2000 and November 2008. The catalog provides also the involved organization, the date and amount of loss (measured as the numbers of ID stolen). Data are homogeneously sampled among various types of organizations: business (35%), education (30%), governments (24%) and medical institutions (10%). We define an event following the procedure described in reference [2,14]. For instance, the largest entries in the data set are (i) the discovery and disclosure of an attack over several years of the TJX Companies with a probable exposition of more than 90 millions IDs (end of the event: January 2007), (ii) the Cardsystems' hack impacting 40 million Visa, MasterCard and American Express cardholders (June 2005), (iii) America Online (30 million credit card ID exposed in 2004), and (iv) the US Department of Veterans Affair (more than 25 million of ID stolen in 2006).

An important issue is that of the reliability and the completeness of the catalog. First, each of the 956 documented events has been confirmed at least by one major media source and/or by an official statement issued by the government. This publicly disclosed information is easily traceable and peer-reviewable (cf. http://datalossdb.org/primary_sources). Note also that www.datalossdb.org (or its predecessor www.attrition.org) is recognized as the most complete dataset by practitioners ranging from computer scientists to lawyers (see e.g. William Roberds and Stacey L. Schreft, Data Breaches and Identity Theft, Federal Reserve of Altanta, Working Paper, 2008, and Jari Raman, Computer Law & Security Report, 2008). This suggests that reported events have rather reliable documented characteristics: the dates of occurrence, the locations and the severity of the events are thoroughly documented by concordant and independent sources. While some reporting errors are bound to exist as in any database, the metrics that we use are generally well-known and disclosed by the organizations themselves, in conformity with the data breach notification act adopted since 2003 by the large majority of states in the US.

Another issue is that of completeness (i.e., unreported events) but, as we discuss below, our analysis suggests that the catalog of the Open Security Foundation provides a reasonable representative sample of the overall activity of ID thefts occurring on the Internet and especially for the US, for the most important events in terms of the number of ID thefts. Furthermore, random errors or censorship in a database generated by a power law distribution do not lead to biases as the empirical distribution converges towards the true distribution. Actually, our analysis suggests a quantification of the degree of incompleteness, which we find all the more acute, the smaller is the event. By proposing a quantitatively falsifiable hypothesis for the distribution of ID thefts, our analysis provides strong incentive for the development of even better databases. As in many other fields of investigation before, it is the never ending back and forth iteration between data base improvements pushed by new insights provided by analysis and the new tests developed by the analysis that pushed knowledge forward.

More deeply, even if the reported events in www.datalossdb.org are traceable or peer-reviewable, it is important to realize that the database construction is based entirely on voluntary efforts, with no forcing function that convinces involved parties to contribute. As a result, one may worry that the observed characteristics might be an artifact of how the database was populated over time and by whom. This issue is actually characteristic of the so-called open-source approach versus the economically-based control approach. The question is to assess the quality and reliability of those voluntary contributions that are intrinsic to the open source approach, as opposed to scientifically organized top-down database and project constructions. It turns out that this question is not specific to www.datalossdb.org but is now investigated in an exploding literature concerned with the pros and cons of open source approaches and with the motivations of open source contributors. Overall, studies have shown that the open source model is more and more endorsed by industry as a viable complementary business approach as well as source of important information [11,12]. We will cite a single representative example for illustration, Wikipedia, a free online encyclopaedia that anyone can edit. A study of 42 entries by Nature magazine on December 14, 2006 put Wikipedia almost on a par with Britannica in terms of accurate science coverage [13]. Nature found that the average science entry in Wikipedia had four errors while Britannica had three.

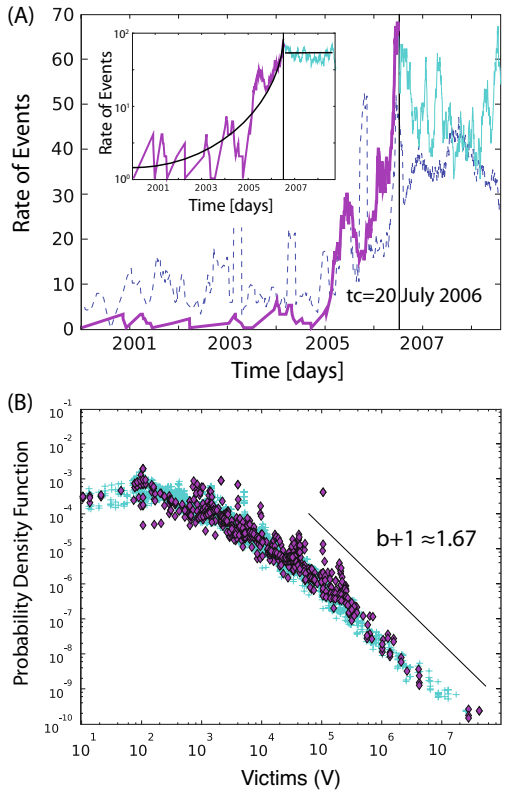

Fig. 1. (Color online) (A) The rate of ID loss events in sliding windows of fifty days is plotted as a function of time, revealing the existence of two successive regimes: (i) explosive growth culminating in July 2006 (red thick line) and (ii) stable rate thereafter (blue thin line). As a matter of comparison, the dashed line represents the rescaled evolution of the rate of new software vulnerabilities. The Spearmann rank correlation between vulnerabilities and ID losses ($\rho = 0.64$) confirms the common trend of cyber risks. The inset shows the plot of the logarithm rate of ID loss events as a function of t. Before the peak the line is upward curved, which confirms an accelerated growth before July 2006. The noisy upward curvature suggests a faster-than-exponential growth before July 2006 (an exponential growh would be qualified by a straight line in the log-lin plot). The black line is a guide showing the upward curvature followed by a plateau. (B) Scatter proxies of probability density functions (PDF) of the size of events obtained in sliding windows of 100 days duration. PDFs obtained by binning or with the adaptive Gaussian kernel density estimator [35] provide similar results. The size of an event is defined as the total number of IDs lost in that event. For the sake of clarity, we show only one PDF out of every fifty PDFs. Red diamonds (respectively blue crosses) correspond to the PDFs obtained before (respectively after) the peak in July 2006.

2.2 Transition from explosive growth to statistical stationarity

The total rate $C(t)$ of ID theft events (measured by the number of events in a sliding window of 50 days) is shown in the top panel of Figure 1 as a function of time. This panel reveals the existence of two distinct phases. Starting from 2000, one can observe a dramatic increase of the rate of attacks up to a peak reached in July 2006, followed by a plateau thereafter. The inset shows a non-parametric evidence suggesting that the first regime was characterized by a strong bursty acceleration, perhaps faster-than-exponential growth as suggested by the upward curvature observed before July 2006 in this linear-logarithmic plot (note that a straight line would qualify an exponential growth). Such singular behavior characterized by a transient explosive growth, which can be mathematically modeled by a power law with finite-time singularity, is often the diagnostic of an impending change of regime [15–17], which we indeed observe beyond the peak in July 2006. It suggests to interpret the time evolution of the rate of ID loss events as first undergoing a non-sustainable growth followed by a maturity period which characterizes the present epoch.

As a verification step supporting our claim that the behavior shown in Figure 1 is not spurious and does not result from reporting biases, we compared normalized ID losses with the time series of new vulnerabilities, which are systematically recorded by the US-CERT (see `http://www.us-cert.gov/cve.html`). The dashed line in Figure 1 shows the rate of recorded software vulnerabilities rescaled so as to be comparable to the ID loss time series. We use the Spearman rank correlation to measure the dependence between the rate of ID losses and that of the vulnerabilities and obtain $\rho = 0.64$. This shows that the dynamics of ID losses is consistent with another coarse grained measure of cyber insecurity. In doing so, we do not claim any direct causality between vulnerabilities and ID losses, but rather propose that the two time series reflect the same underlying growth of cyber-risks.

The lower panel of Figure 1 demonstrates that the distribution pdf(V) of event sizes (defined as the total number of ID stolen in that event) has remained stable, within statistical fluctuations, over the whole time period investigated here from 2000 to Nov. 2008. There is no significant difference between the probability density functions (PDF) in the growth regime before July 2006 (red circles) and during the maturity period (blue diamonds), as evidenced by the perfect collapse of the PDFs. Indeed, Q-Q plots of one sample as a function of other samples and in function of the entire sample, were found to be approximately linear with slope slope $\approx 0.9 \pm 0.3$. This simple non-parametric test is particularly important to ensure a robust interpretation of the data, which is reported according to a "best effort" principle, but without warranty of missed events or inaccuracies, given our remarks in section 2.1. This test supports the robustness of the PDFs of event sizes over time, independently of the event rates. It also contributes to developing some trust in the hypothesis that the PDFs of event sizes have an asymptotic power law shape. This suggests that the mechanism underlying the loss of ID has remained stable, notwithstanding the enormous evolutions that have occurred over this whole time period.

The two pieces of information provided by the two panels of Figure 1 imply that the rate $N(V, t)$ of events of size

V occurring at time t can be factorized under the form

$$N(V,t) = C(t) \cdot \mathrm{pdf}(V), \tag{1}$$

where $C(t)$ and $\mathrm{pdf}(V)$ constitute two independent contributors to cyber-risks. The macro-variable $C(t)$ embodies the overall evolution of the level of threat associated with ID losses. In other words, it provides a metric quantifying the systemic "state of insecurity" of the Internet. In contrast, $\mathrm{pdf}(V)$ measures the relative frequency of large versus small ID losses. While the rate of attacks has varied enormously between 2000 and 2008 as shown by the behavior of $C(t)$ in the upper panel of Figure 1, the relative frequencies of various event sizes has remained remarkably stable, as shown in the lower panel of Figure 1. We now turn to the determination of $\mathrm{pdf}(V)$ in order to characterize quantitatively the level of cyber risks per event.

3 Distribution of ID theft event sizes and consequences

3.1 Power-law versus stretched exponential

Given the result of the previous section that a unique distribution $\mathrm{pdf}(V)$ is sufficient to describe the frequency of event sizes in all time windows from 2000 to 2008, we now determine $\mathrm{pdf}(V)$ by using the largest possible statistical sample including all events of this period. Figure 2 presents the (non-normalized) empirical survival (also called complementary cumulative) distribution function $\bar{F}_u(V)$, defined as the probability that the number of victims in a given event is larger than or equal to V in the range $V \geq u$. Note that $\bar{F}_u(V)$ has a shape similar to the PDFs shown in the lower panel of Figure 1 with an approximately straight tail in this double-logarithmic scale, suggesting a power law distribution

$$\bar{F}_u(V) = \left(\frac{u}{V}\right)^b, \quad \text{for} \quad V \geq u. \tag{2}$$

This power law (2) is observed over more than three decades above the lower threshold $u \approx 7 \times 10^4$. A maximum likelihood estimation (MLE) of the exponent determines $b = 0.7 \pm 0.1$. If model (2) is a correct description of the survival distribution, then $\mathrm{pdf}(V) \sim 1/V^{1+b}$, which is shown as a straight line with slope -1.7 in the lower panel of Figure 1. This result suggests that ID thefts have statistics similar to those observed in the large class of systems with heavy-tails, such as firm and city sizes in the social sciences or earthquakes and other calamities in the natural sciences.

However, visual evidence and MLE are not sufficient to demonstrate that the power law (2) is adequate to describe our statistical data of ID thefts, as discussed in several earlier works [18–20]. To prove that the one-parameter power law (2) is sufficient, we embed it into a broader two-parameter law that have previously been reported to provide a flexible model of many empirical fat-tailed distribution [18] and perform a standard log-likelihood ratio (Wilks) test. Specifically, we use the "stretched exponential" (SE) family

$$\bar{F}_u(V) = \exp\left[-\left(\frac{V}{d}^c\right) + \left(\frac{u}{d}^c\right)\right], \quad \text{for } V \geq u, \tag{3}$$

where c and d are respectively the shape and scale parameters of the SE distribution. Malevergne et al. [20] have shown that the power law family (2) is asymptotically embedded in this SE family in the limit

$$c\left(\frac{u}{d}\right)^c \to b, \text{ as } c \to 0. \tag{4}$$

This has two practical applications: (i) the calibration of c and d for a given u provides an alternative determination (using (4)) of the exponent b of the power law (2) if c is indeed small (typically less than 0.3); (ii) we can use the formal likelihood ratio test of embedded hypotheses which has been shown to hold for the power law seen as asymptotically embedded in the SE family [20,21], to determine whether the one-parameter power law is sufficient or a two-parameter distribution like the SE is necessary. Inset (a) in Figure 2 shows the estimated exponent b obtained from the maximum likelihood estimation (MLE) of c and d translated into b via the equation $b_{\mathrm{SE}} = c(u/d)^c$ derived from (4), as a function of the lower threshold u. For $u \geq 7 \times 10^4$, we obtain an excellent confirmation of the value $b \simeq 0.7 \pm 0.1$ determined from the direct MLE of the power law (2). Inset (b) in Figure 2 shows in addition the logarithm of the likelihood ratio (LLR) of the power law versus the SE fits: for $u < 7 \times 10^4$, LLR < 0 indicating that the power law is not sufficient and that the SE is necessary; in contrast, for $u \geq 7 \times 10^4$, the power law is sufficient and the SE is not necessary, degenerating into the power law as the condition (4) becomes valid.

3.2 Evidence for incompleteness of reported losses for small event sizes

We now discuss two possible hypotheses for the observed cross-over at $u \approx 7 \times 10^4$ below which the distributions shown in the lower panel of Figure 1 and in Figure 2 exhibit a significant downward curvature characterizing a deviation from the power law (2).

A first possible interpretation is that this deviation from the power law reflects the fact that hackers are preferentially targeting large organizations offering substantial potential gains. As a consequence, there would be practically no ID thefts involving only a few individuals. This hypothesis does not stand closer scrutiny: there is strong evidence that millions of home computers are compromised [8] via the use of botnet deployment mechanisms centrally managed by pirates [7], with each computer infection being a unique event potentially leading to ID thefts limited to those IDs which are stored in the computer. According to Vinton Cerf, 100−150 millions computers over a total of 600 millions are compromised [22]. As a rough estimation, assuming that all computers have about the same probability of being infected

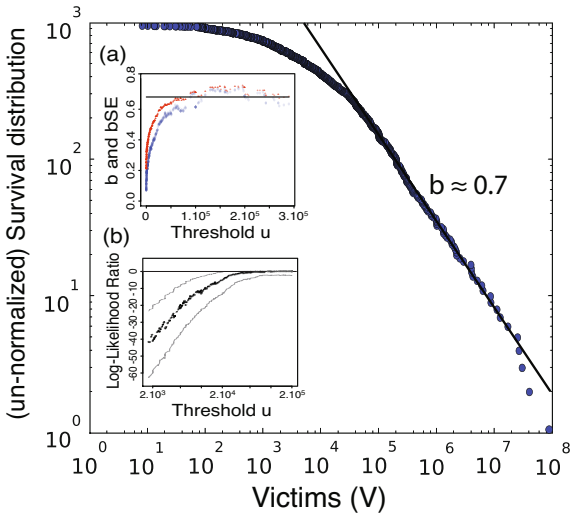

Fig. 2. (Color online) Non-normalized survival distribution (double logarithmic scale) of ID losses, constructed using the data provided in [2] the straight black line is the fit with the power law (2) with $b = 0.7$ for number of victims larger that the lower threshold $u = 7 \times 10^4$. The red dashed line is the fit with the Stretched Exponential (SE) defined by expression (3). Inset (A) shows the dependence of the index b as a function of u obtained directly from the maximum likelihood estimation (MLE) of the exponent of the power law (2) (crosses) and indirectly from the MLE of the parameters c, d of the stretched exponential (SE) law (3) using the correspondence $b_{\rm SE} = c(u/d)^c$ (diamonds) as described in the text. The horizontal line is at $b = 0.68$. Inset (B) shows the logarithm of the likelihood ratio (LLR) of the power law versus the SE fits, which converges to 0 as u increases, thus demonstrating that the simple one-parameter power law is sufficient and the two-parameter SE law is not necessary to explain the tail of the data set. The two grey lines delineate the 95% confidence interval obtained by bootstrap.

and counting one computer per Internet user, this implies that about one sixth of US computers are exposed. Thus, about 50 millions US citizen are constantly exposed to attacks targeting their own computer. Such events should thus provide a huge population of small ID theft events' which is absent from even the most complete dataset of the Open Security Foundation [2].

3.3 Super-linear growth of the ID loss threat

There is another remarkable consequence deriving straightforwardly from the power law (2) with exponent $b < 1$. Indeed, the smallness of the power law exponent $b < 1$ implies a typical faster-than-linear growth of cumulative losses with time. Because $b < 1$ and since (i) there is no upper threshold yet relevant and (ii) the lower threshold $u \approx 7 \times 10^4$ remains stable over time, the mean and variance of the number of ID losses per event are mathematically infinite. In practice, this means that their values in any finite catalog exhibit growing random fluctuations as the number of recorded events increases, due to the never decreasing influence of the largest event sizes. Then, the cumulative sum $\mathcal{V}(t)$ of all losses over all events up to time t is controlled by the few largest events in the catalog [23]. This leads to a faster-than-linear growth

$$\mathcal{V}(t) \sim t^{1/b} \approx t^{1.4}. \tag{5}$$

This results is solely due to the statistical mechanism that, as more events occur, some are bound to explore more and more the tail of the heavy-tailed power law distribution (2). Note this law (5) constitutes a lower bound, which is attained only when the rate of event occurrences is itself not growing, as seems to be the case since July 2006.

Such faster-than-linear growths due to the pure statistical power law mechanism have been documented in natural hazards for losses caused by floods [24] and for the cumulative seismic energy released at regional scales [25] (see [23] for a detailed mathematical derivation and discussion). Given the heavy-tail nature of the distribution of ID theft numbers per event, we should not be surprised that the Internet appears more and more insecure and dangerous, just as a result of this mechanism.

4 In cyber-risks, size matters

4.1 Cross-sectional universality of ID losses

We have shown that the PDF of event sizes is constant over time. We now investigate whether there exists some difference between the PDFs of event sizes in a cross-sectional analysis of different sectors of activity, which could reveal different vulnerability characteristics.

Our datasource uses four distinct sectors of activity: publicly traded companies (Biz), schools and universities (Edu), governmental agencies (Gov), and medical services (Med). Distinct regulations and industry benchmarking imply that organizations implement homogenous security processes in a given sector, but these security processes operating in a given sector are different from those in a different sector. A priori, one could expect that distinct factors acting in these different sectors imply dissimilar attractiveness to hackers leading to different levels of vulnerability, which should be revealed in the statistical properties of the catalogs of ID losses. In contradiction with this anticipation, the top panel of Figure 3 shows that one cannot reject the hypothesis that the PDFs of ID loss size per event are identical for the four sectors Biz, Edu, Gov, Med.

If two typical organizations belonging to two different sectors are subjected to distinct exposition and permeability threats, the remarkable conclusion suggested by the top panel of Figure 3 is that the associated level of security just compensates for the increasing threat, putting all organizations at a similar overall risk level. This result is reminiscent of the effect documented in references [26,27], that systems exposed to different distributions of attacks converge to similar level of vulnerabilities when they try to optimize their efficiency in the presence of constraints.

This could mean that organizations, which are indeed attempting to optimize their defenses against cyber-risks, may have already reached an intrinsic barrier. With the evolving nature of the threats and given the complexity of the associated processes in the presence of limited resources, the observed level of ID losses may be a robust dynamical equilibrium that will be difficult to improve upon. This suggests that, in absence of a fundamentally new qualitative paradigm, these cyber-risks are bound to remain with us for the foreseeable future.

4.2 Size effects of vulnerabilities to cyber-risks

The bottom panel of Figure 3 plots the PDFs of victims per event sorted by target organization sizes. There are several possible measures for the size of an organization. Here, we take the number of employees, which correlated well with other measures [28]. The PDFs are constructed for 269 universities [29] and 105 publicly traded companies [30]. The good collapse of the PDFs confirms the universality of the power law distribution of event loss sizes, as in Figures 1 and 2.

However, the tails of the PDFs are truncated at upper values which seem to grow with the organization sizes. This size effect is better revealed by the scatter plot of the inset in the bottom panel of Figure 3, which shows that the largest losses $V_{\max}$ for a given range of organization sizes S seem to grow with S. This visual impression is confirmed by performing linear regressions of $\log V(q)$ as a function of $\log S$, $\log V(q) = \sigma \log S + \epsilon$, where $V(q)$ is the 99% quantile of the losses for a given organization size S. We find a stable determination of the exponent $\sigma \approx 1.3 \pm 0.1$. This means that the largest losses for a given set of entities of size S grow with S as $V_{\max} \sim S^{\sigma} \approx S^{1.3}$.

Naively, one would have expected a linear growth with $\sigma = 1$. The faster-than-linear law may express a combination of effects, which include a faster-than-linear growth of the number of IDs stored in a given entity as a function of its number of employees, a bigger exposition that makes the attacks of large entities more attractive to hackers and possibly a greater vulnerability due to more bridges or "boundaries" with the external world which are more difficult to manage. The faster-than-linear law is characteristic of a size effect which is similar to the size effects documented for instance in material failure [31] and species fragility [32].

We now show how σ is related to the exponent b of the PDFs of event loss sizes defined in (2). For this, we write the probability $\Pr(\text{ID losses} \geq V)$ to find an event with more than V ID losses as

$$\Pr(\text{ID losses} \geq V) = \int_{S_{\min}}^{+\infty} dS Z(S) \Pr_1(\text{ID losses} \geq V|S), \tag{6}$$

where $S_{\min}$ is a minimum size for an organization to be viable, and $Z(S)$ is the distribution of organization sizes, well-known to follow Zipf's law ($Z(S) \sim 1/S^{1+\mu}$ with $\mu \approx 1$) [28,33,34] so that $Z(S)dS$ is the number of organizations with sizes between S and $S + dS$. Moreover,

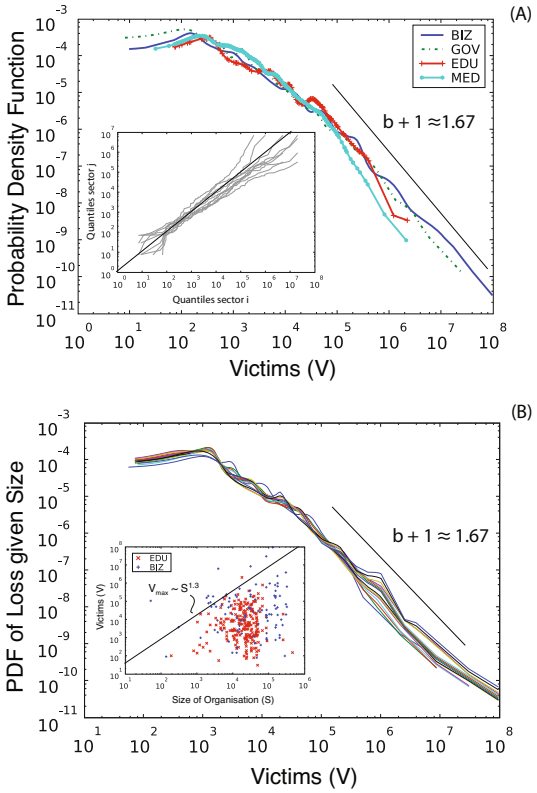

Fig. 3. (Color online) (A) Probability density functions of the number of victims (V) per event sorted by sector: business (Biz), governmental agencies (Gov), schools and universities (Edu), medical industries (Med). Inset shows quantile-quantile plot (with 5% interquantiles) of sectors taken against each other. Linear fit obtained for the presented lines show that we cannot reject that *slope* $= 1$, ruling out the hypothesis that distributions are different. (B) Probability density functions (PDF) of victims per event sorted by sizes of the target organizations. We construct one PDF per decade in organization sizes, i.e., we collect all events occurring for organizations of sizes between S^* and $10 \times S^*$ and construct the corresponding PDF. We then vary S^* across the whole sample (to avoid overlapping we take only one out of fifty PDFs). All PDFs exhibit a good collapse, confirming the universality of the power law distribution of event loss sizes, as in Figures 1 and 2. Similarly to presented above, by performing linear regressions of (log) quantiles of all samples, we cannot rule out that all samples are drawn from the same probability distribution. The inset shows in double logarithmic scale a scatter plot of the losses (V) as a function of size for 374 entities. The straight line with slope $\simeq 1.3$ is the best linear fit ($p = 0.00$ and $R^2 = 0.74$) of the 99% percentile of the logarithmic losses for both 269 universities (blue plus symbols) [29] and 105 publicly traded companies (red crosses) [30] as a function of organization logarithmic size.

$\Pr_1(\text{ID losses} \geq V|S)$ is the probability to find an event with more than V ID losses in a given organization of size S. We know one property of $\Pr_1(\text{ID losses} \geq V|S)$, namely that it drops abruptly to vanishing values for $V > CS^{\sigma}$, where C is a positive constant, as documented above. This implies that, for a fixed V, all integrants with $S < (V/C)^{1/\sigma}$ do not contribute to the integral. Motivated

by the power law (2), we also assume a power law shape for $\Pr_1(\text{ID losses} \geq V|S)$ with exponent b_1. Putting all this together, expression (6) becomes

$$\Pr(\text{ID losses} \geq V) \simeq \int_{S_{\min}(V)}^{+\infty} \frac{dS}{S^{1+\mu}} \frac{1}{S^{b_1}}, \qquad (7)$$

with $S_{\min}(V) \sim (V/C)^{1/\sigma}$. This yields $\Pr(\text{ID losses} \geq V) \sim 1/S^{b_1+\frac{1}{\sigma}+(\mu-1)}$. Identifying this power law with (2) in the tail gives $b = b_1 + \frac{1}{\sigma} + (\mu - 1)$. Given that $\sigma \approx 1.3 \pm 0.1$, we have $1/\sigma \approx 0.77 \pm 0.1$. Since $b = 0.7 \pm 0.1$, this calculation allows us to infer that the distribution of ID losses for a given organization is fairly flat ($b_1 \simeq 0$). In other words, the efforts necessary to get just a few or a large number of IDs are not much different, once an organization has been compromised. Our conclusion does not rely sensitively on the validity of Zipf's law. However, the value $b < 1$ imposes a bound on the exponent μ of Zipf's law which cannot be significantly larger than 1.

5 Conclusion

We have presented three different tests that confirm the general validity and robustness of the probability distribution of ID losses per event (where ID losses has been taken as a proxy for information risks on the Internet). We showed that the PDFs are the same irrespective of (i) the growth phase before July 2006 versus stationary regime thereafter, (ii) the sectors of activity, and (iii) the size of targeted organisations. Returning to the questions raised in the introduction, it is striking and a priori counter intuitive to find that all organisations are evenly vulnerable, whatever their implemented information security. This raises important questions concerning the tradeoff between exposition and counter-measures in the complex evolving landscape of cyber-risks. The consequences on the evolution of the Internet remain to be studied. This present paper provides a first partial approach of the study of the development of the Internet and of cyber-risks taking into account their intricate entanglement.

We have shown the existence of a size effect, such that the largest possible ID losses per event grow faster-than-linearly with the organization size. This has led us to derive two important consequences. First, the small value $b \simeq 0.7$ of the power law distribution of ID thefts is explained by interplay between Zipf's law and the size effect. Second, we have found indirect evidence that compromised entities typically expose to hackers a small or large number of IDs with basically the same frequency. This inference is very important for the quantification of cyber risks and suggests that counter-measures should be targeted towards building internal barriers, avoiding the "Titanic" effect of inadequate compartmentalization.

A limit of our study is that we have analyzed only one class of cyber risks (ID thefts) while the subject is much richer, including defacing home pages, hacking, malicious code (such as viruses and worms), denial of service attacks, theft of information via e.g. phishing and other means, fraud, corruption of data, insider exploitation and even cyber terrorism. However, we believe that our conclusions are relevant and useful for several reasons. First, identity theft is soaring and is in fact the fastest growing crime in the US, according to the Federal Trade Commission. Tammy J. McInturff, Technology Editor at LOMA (an international association of more than 1200 insurance and financial services companies from over 80 countries, http://www.loma.org) reports that "ID theft is far greater in terms of damages to business consumers than many people actually think. Identity theft has risen rapidly because companies have moved to larger and larger acquisition of consumer data. This has made it easier for a thief to not only steal one credit card number at a time, but also 300 000 others at the same time". There are hundreds of articles in the literature (a few are quoted here), explaining that identity or simply data thefts is today by far the main risk on the Internet. IT Security Experts agree that stealing data is the common denominator of the vast majority of attacks, because they are concealable. Here it is important to stress that we do not make any assumptions on the types of attacks (which are indeed numerous). Finally, we end with an hypothesis, open to falsification as more data become available, that other types of cyber risks will exhibit similar statistical characteristics. This paper has opened a window to an important societal question, which ought to be addressed within a systematic scientific methodology, without waiting for more data but pushing by its challenging conclusions for more and better data.

We are grateful to Stefan Frei for fruitful discussion regarding the dependence between vulnerabilities and ID losses and for providing the vulnerability time series. This work was supported by the Swiss National Foundation, grant 2-77059-07 and by the MTEC Foundation (Fordergesellschaft fur Betriebswissenschaften MTEC), We also acknowledge financial support from the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01-08-2.

References

1. G. Hamilton, Viruses, New Scientist **2671**, 38 (2008)
2. datalossdb, http://datalossdb.org/(06.01.2009)
3. S. Frei, M. May, U. Fiedler, B. Plattner, Large Scale Vulnerability Analysis, ACM SIGCOMM 2006 Workshop, 2006
4. J. Zittrain, *The Future of the Internet–And How to Stop It*? (Yale University Press, 2008)
5. R. Mansell, B.-S. Collins, *Trust and crime in information societies* (Edward Elgar Northampton, MA, 2005)
6. K. Anderson, E. Durbin, M. Salinger, J. Econ. Perspect. **2**, 171 (2008)
7. D. Dagon et al., 23rd Annual Computer Security Applications Conference (2007)
8. United States Attorney's Office, http://www.usdoj.gov/usao/cac/pressroom/pr2007/143.html(06.01.2009)
9. B. Schneier, Risks of third-party data, Communications of the ACM **48**, (2005)

10. B. Koops, R. Leenes, Datenschutz und Datensicherheit-DuD **30**, 553 (2006)
11. von Krogh, G.S. Spaeth, K.R. Lakhani, Research Policy **32**, 1217 (2003)
12. S. Haefliger, von Krogh, G.S. Spaeth, Management Science **54**, 180 (2008)
13. J. Giles, Nature **438**, 900 (2005)
14. R. Hasan, W. Yurcik, A statistical analysis of disclosed storage security breaches, *Proceedings of the second ACM workshop on Storage security and survivability* (2006), pp. 1–8
15. A. Johansen, D. Sornette, Physica A **294**, 465 (2001)
16. K. Ide, D. Sornette, Physica A **307**, 63 (2002)
17. S. Gluzman, D. Sornette, Phys. Rev. E **6601**, 016134 (2002), N1 PT2:U315-U328
18. J. Laherrère, D. Sornette, Eur. Phys. J. B **2**, 525 (1998)
19. A. Clauset, C.-R. Shalizi, M.-E.-J. Newman, Power-law distributions in empirical data (2007), http://arxiv.org/abs/0706.1062
20. Y. Malevergne, V.F. Pisarenko, D. Sornette, Quantitative Finance **5**, 379 (2005)
21. Y. Malevergne, D. Sornette, *Extreme Financial Risks: From Dependence to Risk Management* (Springer, Heidelberg, 2006)
22. BBC article on Vinton Cerf's WEF talk, http://news.bbc.co.uk/2/hi/business/6298641.stm(06.01.2009)
23. D. Sornette, *Critical Phenomena in Natural Sciences*, 2nd edn. (Springer Series in Synergetics, Heidelberg, 2006)
24. V. Pisarenko, Hydrol. Proc. **12**, 461 (1998)
25. M. Rodkin, V. Pisarenko, Comput. Seismd. **2000**, 242 (2001) (in Russian); English translation in Computational Seis. Geodyn. **5**
26. J.M. Carlson, J. Doyle, Phys. Rev. Lett. **84**, 2529 (2000)
27. J. Doyle et al., Proceedings of the National Academy of Sciences **41**, 14497 (2005)
28. R. Axtell, Science **293**, 1818 (2001)
29. Source: Official websites from universities. University population is composed with students, graduates, administrative staff and faculty
30. Source: Bloomberg. Number of employees of publicly traded companies exposed to ID loss at the time of the event
31. Z. Bažant, Int. J. Frac. **83**, 19 (1997)
32. M. Cardillo, Anim. Cons. **6**, 63 (2003)
33. G. Zipf, *Human behavior and the principle of least effort* (1949)
34. X. Gabaix, The Q. J. Econ. **114**, 739 (1999)
35. K. Cranmer, Computer Physics Communications **3**, 198 (2001)